MULTIMODAL VIDEO REPRESENTATION LEARNING WITH CONVOLUTIONAL NEURAL NETWORKS

Cheng Wang, Haojin Yang, Christoph Meinel

Hasso Plattner Institute, University of Potsdam
Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam, Germany
Email:{cheng.wang,haojin.yang,christoph.meinel}@hpi.de

ABSTRACT

In this paper we present a framework for multimodal video representation learning. Our framework takes advantage of spatial, temporal and auditory modalities, and fuses multimodal features intended to create more robust video representation. The Convolutional Neural Network (*CNN*) models at the center of this work are trained for each modality respectively. We further fuse the multimodal features by training a Deep Neural Network (DNN) on the top of three modalities. The evaluation results show that for video classification task various modalities are complementary to each other. Furthermore, we conducted in-depth study for measuring the reasonable experimental options that may affect the classification performance on multimodal as well as unimodal resources.

Index Terms— video representation, multimodal feature fusion, convolutional neural network

1. INTRODUCTION

Nowadays, automatic video classification becomes a more desirable technique due to the rapid increment of web video data. Video classification is a task that heavily relies on the ability of feature representations. Recent work [1, 2] show that robust feature can be learned with deep neural network from spatial and temporal stream respectively. It significantly outperforms hand-crafted shallow feature such as SIFT [3] for image representation and dense trajectories [4] for capturing temporal information. In [1, 2], it has been proven that the video classification accuracy can be largely improved by combining multimodal information (spatial and temporal modalities) comparing to unimodal based approach [5, 6, 7]. This result confirms the finding of [8, 9], which pointed out that exploring the relationships of different modalities is beneficial for classification task. However, those recent work do not take advantage of audio information, which is also one of the most important features for video representation. Therefore in this work, we assume that the audio stream is a useful complementary feature, and propose to combine spatial, temporal and acoustical clues with CNNs to learn unified video representation. The main contribution of this work can be summarized as follows:

(1) We proposed an mutli-stream deep learning framework based on convolutional network for video classification.

(2) We investigated the possibility of fusing spatial, temporal and audio features, and proved the feasibility.

(3) Our extensive experiments show that audio information can be beneficial in improving accuracy from unimodal feature. Our fusion result of spatial and temporal modality achieves competitive result to the state-of-the-art.

2. METHODOLOGY

2.1. Overall Architecture

Figure 1 presents the overall architecture of our framework. For each modality, we first learn the discriminative features with CNN separately. Then we extract features from the sixth fully connected layer (fc-6) by using pre-trained CNN models. In order to explore the non-linear dependencies between spatial, temporal and acoustical CNN features, we propose to add an additional fusion layer (fully connected layers or SVM (support vector machine) classifier) at the end of the network, and to learn the fused features. The output will then feed into a softmax classifier to obtain the prediction result.

2.2. Spatial and Temporal CNN

Mimicking the human vision mechanism, the video data can be decomposed into sequential image frames. The most important clues here are the visual presentation of each single frame and the consecutive changes between them. Therefore, the visual representation of video data can be briefly separated into spatial and temporal modalities. Spatial information, as e.g., object shape, color, contrast, brightness help human visual system to recognize objects, visual changes from static images. The complementary temporal component contains motion information across video frames.

Recently deep neural networks achieved promising results in many vision research fields, it shown superior ability

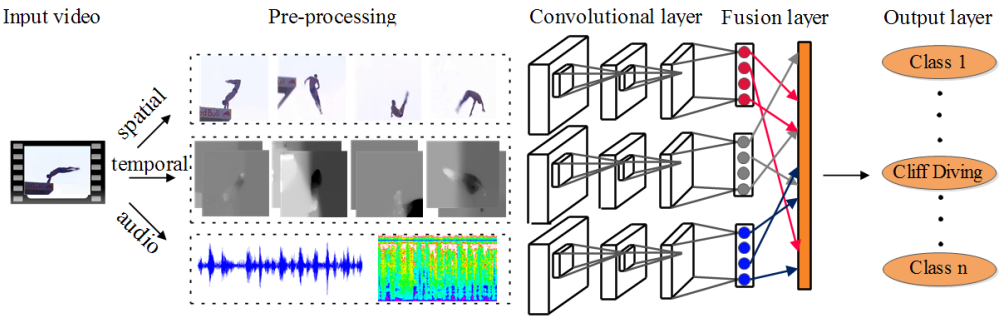


Fig. 1: Framework for multimodal video representation. First, we train deep models from spatial, temporal and acoustical clues separately. Second, we impose fully connected layers or SVM classifier as fusion layer to learn the non-linear dependencies among different modalities, and combine them as unified video representation. The inputs are keyframes extracted from video for spatial CNN, X-direction and Y-direction optical flow images for temporal CNN, and frequency spectrum images extracted from raw audio waveform for acoustical CNN, respectively. All input images have dimension of 224×224.

for solving visual classification problems[10, 11]. In [1], Simonyan et al. investigated the possibility of building spatial and temporal CNN models for video classification. Inspired by this work we also trained spatial and temporal CNN models for the visual video representation.

The spacial CNN is trained on the still images extracted from video with proper frame rate. To learn a deep neural network large amount of training data is usually required, since there might be several million of parameters need to be optimized. Lacking of training data becomes a general issue in deep learning applications. However, base on a small dataset, it is still possible to obtain an acceptable result by using fine-tuning technique, by which a robust deep model will be used for offering optimized initial weights. The experimental results of [1] also confirms this assumption. Since the applied UCF-101 dataset is relative small, the accuracy of from scratch trained model is about 20% lower than the best fine-tuned model. In our framework, we applied a very recent deeper network VGG_19 [12] to fine-tune our spacial model, which significantly extends the depth of the network. As described by its name, VGG_19 consists of 19 layers including 16 convolutional layers and 3 fully connected layers. All convolutional filters decreases to 3×3 and the stride reduces to 1, which is intended to let the network to learn with more subtle granularity. The reported top-5 error rate of this model on the ILSVRC-2012 validation set is 7.0%. The detailed parameter configuration will be discussed in Section 3.

The temporal CNN is intended to capture the consecutive changes of video frames. To this end optical flow is applied as the input for training the temporal CNN. We consider an optical flow as a set of displacement vector fields d_t between the sequential frame pairs t and $t+1$, where t is a time point within video. The horizontal and vertical flow vector d_t^x and d_t^y are created independently, which could be seen as two image channel types. Thus, for each extracted video frame at time t, we create optical flow images based on its L subsequent frames, and stack them into a new image with $2L$ channels. This way within each newly created flow image a consecutive motion scene has been stored and the scene frames share one event label. We applied the VGG_M [12] architecture to train the temporal CNN which is based on 5 convolutional layers and 3 fully connected layers

2.3. Acoustical CNN

Audio soundtracks are important clues for video understanding and could serve as complementary data for other modalities [8, 13]. Therefore in this work we intended to investigate the influence to the video classification performance by adding an acoustical CNN model. To train the acoustical CNN, we consider two types of auditory inputs including Mel-Frequency Cepstral Coefficients (*MFCC*) and audio spectrogram. We conducted extensive experiments on both feature types to observe their performance on video classification task.

First, we applied a 3-layer neural network to train audio model with MFCC features. To apply the audio spectrogram, we transformed 1-D soundtrack to 2-D spectrogram images, and then train acoustical CNN with them. We also adopted VGG_M, VGG_19 models to fine-tune our acoustical CNN, based on that we extracted spectrogram CNN features from the fc-6 layer, and further applied in the fusion stage.

2.4. Fusion Layer

To fuse the features extracted from convolutional networks, we adopted two commonly used fusion approaches: Early Fusion (EF) and Late Fusion (LF) [14] on neural network (NN) and support vector machine (SVM) respectively. In EF, all features are averaged and concatenated into a vector for training classifier. In LF, the prediction scores are combined as the training input.

Given N video training samples, each of which is represented as: {spatial feature, temporal feature, acoustical feature, ground truth label} after feature extraction. They can be denoted as D=$\{x_s^{(n)}, x_t^{(n)}, x_a^{(n)}, y^{(n)}\}_{n=1}^{N}$. Each input vector has the same dimensionality M and represented as x_s=$\{x_s\}_{m=1}^{M}$, x_t=$\{x_t\}_{m=1}^{M}$, x_a=$\{x_a\}_{m=1}^{M}$ and y=$\{y_k\}_{k=1}^{K}$.

In NN, lets assume the input vector at layer l-1 as x_i, the pre-activation value j^{th} unit at l layer can be formulated as

$$z_j^{(l)} = \sum_{j=1}^{N^{(l-1)}} W_{ij}^{(l-1)} x_i + b_i^{(l-1)} \quad (1)$$

where $N^{(l-1)}$ represents the number of units at l-1 layer. Then the activation of j^{th} neural unit at layer l is computed by $\hat{y}_j^{(l)} = f^{(l)}(z_i^{(l)})(l > 1)$, where $f(\cdot)$ denotes the activation function. In this work, sigmoid function $f^{(l)}(x) = \frac{1}{(1+e^{-x})}$ is used as activation function for all hidden layers, and softmax function is used for activating output layer, that is, $f^{(l)}(x) = \frac{e^{(x-\varepsilon)}}{\sum_{k=1}^{K} e^{(x_k-\varepsilon)}}$, where $\varepsilon = max(x_k)$. The objective function is

$$\underset{\boldsymbol{W},\boldsymbol{b}}{\arg\min}\ C = \frac{1}{2N} \sum_{n=1}^{N} \parallel \hat{y}_m^{(n)} - y^{(n)} \parallel^2 + \frac{\lambda}{2} \sum_{l=1}^{L-1} \parallel \boldsymbol{W_l} \parallel_F^2 \quad (2)$$

where the second term is weight decay for preventing overfitting.

In SVM, we adopted "one-vs-all" classification and χ^2 as kernel, which is computed as $k(x_i, x_j)$=$e^{-\frac{d\chi^2(x_i,x_j)}{\eta}}$, where χ is the mean value of all pairwise distance in training set.

3. EXPERIMENT

3.1. Dataset Description and Preprocessing

We evaluated our approach on UCF-101 dataset [15]. It consists of 13220 video clips cover 101 action categories. All the experiments in this work are based on the first train/test split, namely 9537 clips for training and 3783 clips for test. For simplification the whole dataset is referred to *UCF-101-full*. In the acoustical CNN training, we removed video clips without or with a invalid soundtrack. We finally obtained 6624 clips (4733/1891 for training/test) which fall into 50 categories. Therefore, we conducted all the audio modality involved experiments only with 50 classes, it will be subsequently referred to *UCF-101-50*.

In temporal CNN training, as already mentioned we created stacked flow images for each video frame with $L = 10$, i.e. each stacked flow image has 20 channels. We adopted the off-the-shelf GPU implementation of [16] from the OpenCV[1] library for flow image creation.

In MFCC based acoustical model training, we computed MFCC feature for 20ms time-window with 50% overlap using HTK toolbox[2]. This model is trained by employing a 3-layer neural network. The input for the network is video-level 20-D MFCC feature vector. In the spectrogram CNN training, we transformed the video soundtrack to 2-D spectrogram images as the input with a fixed size of 224×224.

Table 1: Result of spatial and temporal CNN (*UCF-101-full*)

	Description	Accuracy
Spatial	fine-tune last 5 layers	0.753
	train 5-layer DNN on fc-6 output	**0.786**
Temporal	applied VGG_M architecture	0.664
	train 5-layer DNN on fc-6 output	**0.768**
Fusion	LF with SVM	0.822
	EF with 5-layer DNN	**0.851**
Comparison	Karpathy et al. [7]	0.654
	Donahue et al. [2]	0.829
	Wu et al. [18]	0.8416
	Zheng et al. [19]	0.8419
	Srivastava et al. [20]	0.843
	Simonyan et al. [1]	**0.88**

3.2. Results and Discussion

This section give a discussion on our experimental results. We applied Caffe framework [17] for training all CNN models.

3.2.1. Spatial and Temporal CNN

Table 1 shows the top evaluation results of our spatial, temporal and fusion experiments on *UCF-101-full*. The spatial CNN is fine-tuned on the VGG_19 model, where the best result is achieved by training the last 3 fully connected layers and last 2 convolutional layers, while fine-tune more layers didn't increased the accuracy. We further trained a 5-layer DNN (4096/2048/1024/512/101) based on the output of the fc-6 layer of spatial and temporal CNN model respectively. This significantly improved the accuracy. The selection of fc-6 layer is motivated by the experimental results of [2]. Moreover we fused the fc-6 output of spatial and temporal CNN by using linear combination (8192-D), and trained a 5-layer DNN which achieved our best result **0.851**. We also evaluated LF approaches by applying various classifiers, from which SVM achieved the best result. Our DNN based EF achieves competitive result to the state-of-the-art work.

3.2.2. Acoustical CNN

Figure 2 shows the test accuracy of trained acoustical models. We can see that the model trained on MFCC feature performs poor and achieved only 4.7% test accuracy. To train the spectrogram CNN, we first utilized Caffenet architecture

[1] http://opencv.org/

[2] http://htk.eng.cam.ac.uk/

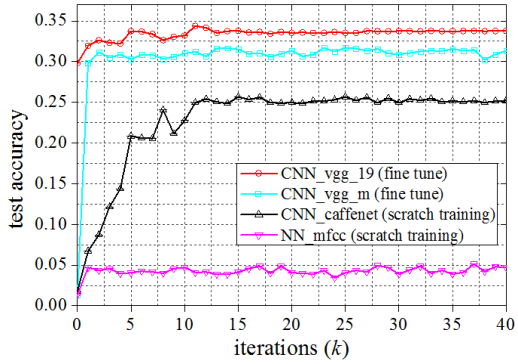
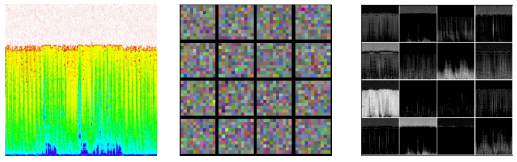


Fig. 2: Acoustical model training (*UCF-101-50*)



(a) input (b) conv 1 filters (c) conv 1 outputs

Fig. 3: Visualization of the first convolutional layer of CNN_caffenet model. We displayed only the first 16 filters.

(contains 5 convolutional layers and 3 fully connected layers) and performed the training from scratch. The achieved accuracy is 25.2%. Further improvements were made by fine-tuning on VGG_M (31.19%) and VGG_19 (33.7%). Therefore, we adopted the fine-tuned model for our fusion experiments. Figure 3 displays the filters and outputs of the first convolutional layer of CNN_caffenet model.

3.2.3. Multimodal Fusion

Our fusion experiments were conducted on Window 8, Intel 3.20GHz×4 CPU and 8G RAM with Matlab. We adopted one hidden layer with 4096 neurons for each NN based fusion, the learning rate is set to 0.001, and momentum=0.095. According to the scale of training set, training batch size is set to 52, epoch number is fixed at 400, weight decay λ=1 × 10^{-4}.

Figure 4 presents fusion results based on features extracted from different modalities. From which we can draw following conclusions: (1) EF-NN and LF-SVM perform better in all modality combinations than EF-SVM and LF-NN. (2) Generally, multimodal fusion can improve the classification result compare to unimodal based approach. (3) By considering audio information, the performance can only be slightly enhanced on *UCF-101-50*. This is due to the fact that many audio tracks are quite noisy or very short, a large part of them only contain background sounds that are unrelated to the video content. Those reasons make it difficult to train an accurate audio model on UCF-101 dataset. Therefore the trained audio model (with only 33.7% accuracy) doesn't retain enough information for complementing other modalities.

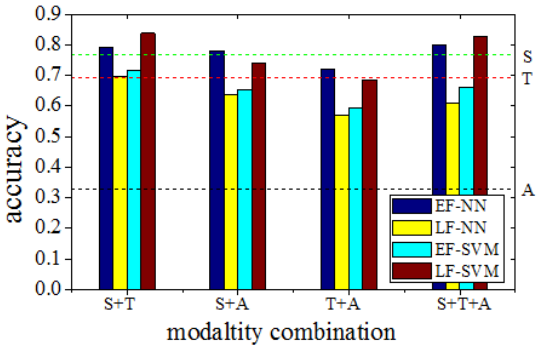


Fig. 4: Fusion result (*UCF-101-50*). S:spatial, T:temporal, A:audio. The right side presents the unimodal accuracy.

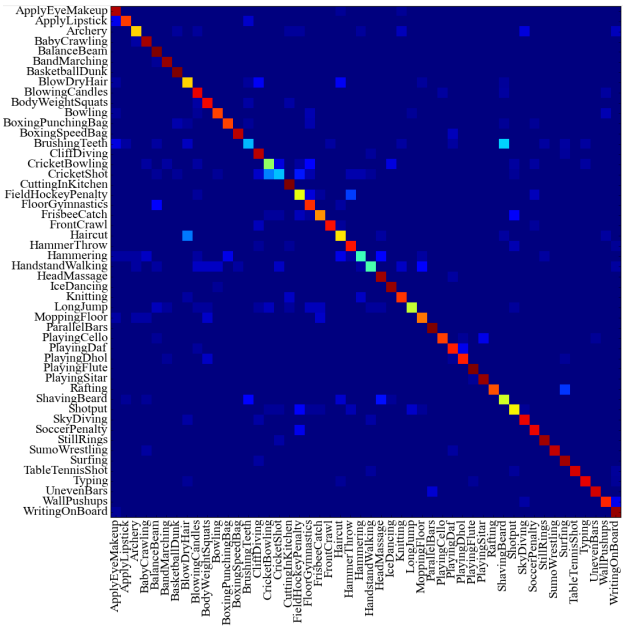


Fig. 5: Confusion matrix of LF_SVM (*UCF-101-50*)

Figure 5 presents the confusion matrix of fusing spatial, temporal and audio modalities.

4. CONCLUSION

This paper has presented a framework for learning multimodal video representation from spatial, temporal and auditory clues. We investigated audio spectrogram images to train acoustical CNN, which is complementary to other modalities. A fusion layer is deployed to combine features that extracted from spatial, temporal and acoustical CNN respectively. The evaluation results show that our proposed DNN fusion network achieved superior result comparing to unimodal methods. As the future work, we will explore more efficient acoustical model and multimodal fusion strategies.

5. REFERENCES

[1] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, 2014, pp. 568–576.

[2] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *arXiv preprint arXiv:1411.4389*, 2014.

[3] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[4] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu, "Action recognition by dense trajectories," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3169–3176.

[5] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu, "3d convolutional neural networks for human action recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 221–231, 2013.

[6] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt, "Spatio-temporal convolutional sparse auto-encoder for sequence classification.," in *BMVC*, 2012, pp. 1–12.

[7] Andrej Karpathy, George Toderici, Sachin Shetty, Tommy Leung, Rahul Sukthankar, and Li Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1725–1732.

[8] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[9] Nitish Srivastava and Ruslan R Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in neural information processing systems*, 2012, pp. 2222–2230.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[11] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014*, pp. 818–833. Springer, 2014.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[13] Zuxuan Wu, Yu-Gang Jiang, Jun Wang, Jian Pu, and Xiangyang Xue, "Exploring inter-feature and inter-class relationships with deep neural networks for video classification," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 167–176.

[14] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.

[15] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[16] Thomas Brox, Andrs Bruhn, Nils Papenberg, and Joachim Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Proceedings of ECCV 2004*, vol. 3024, pp. 25–36.

[17] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[18] Jianxin Wu, Yu Zhang, and Weiyao Lin, "Towards good practices for action video encoding," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 2577–2584.

[19] Jingjing Zheng, Zhuolin Jiang, Rama Chellappa, and Jonathon P Phillips, "Submodular attribute selection for action recognition in video," in *Advances in Neural Information Processing Systems*, 2014, pp. 1341–1349.

[20] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov, "Unsupervised learning of video representations using lstms," *CoRR*, vol. abs/1502.04681, 2015.